Genetics 540
Models of DNA evolution, part 1
Joe Felsenstein
Department of Genetics
University of Washington
joe@genetics
Change of DNA sequence of a species

... is actually mutation, then substitution, in a population
The Jukes-Cantor model (1969)
the simplest symmetrical model of DNA evolution
Transition probabilities under the Jukes-Cantor model

- All sites change independently
- All sites have the same stochastic process working at them
- Make up a fictional kind of event, such that when it happens the site changes to one of the 4 bases chosen at random (equiprobably)
- Assertion: Having these events occur at rate $\frac{4}{3}u$ is the same as having the Jukes-Cantor model events occur at rate $u$
- The probability of none of these fictional events happens in time $t$ is $\exp(-\frac{4}{3}ut)$
- No matter how many of these fictional events occur, provided it is not zero, the chance of ending up at a particular base is $\frac{1}{4}$.
- Putting all this together, the probability of changing to C, given the site is currently at A, in time $t$ is
  \[
  \text{Prob} (C|A, t) = \frac{1}{4} \left( 1 - e^{-\frac{4}{3}ut} \right)
  \]
  while
  \[
  \text{Prob} (A|A, t) = e^{-\frac{4}{3}t} + \frac{1}{4} \left( 1 - e^{-\frac{4}{3}ut} \right)
  \]
  or
  \[
  \text{Prob} (A|A, t) = \frac{1}{4} \left( 1 + 3e^{-\frac{4}{3}ut} \right)
  \]
  so that the total probability of change is
  \[
  \text{Prob} (\text{change}|t) = \frac{3}{4} \left( 1 - e^{-\frac{4}{3}ut} \right)
  \]
The fraction of sites differing after branches of different length, under the Jukes-Cantor model.
Kimura’s (1980) K2P model of DNA change, which allows for different rates of transitions and transversions,
A similar derivation can be done for the K2P model. It involves making two kinds of fictional events:

- I. At rate $\alpha$, if the site has a purine (A or G), choose one of the two purines at random and change to it. If the site has a pyrimidine (C or T), choose one of the pyrimidines at random and change to it.
- II. At rate $\beta$, choose one of the 4 bases at random and change to it.

By proper choice of $\alpha$ and $\beta$ one can achieve the overall rate of change and $T_s/T_n$ ratio $R$ you want. For rate of change 1, the transition probabilities (*warning: terminological tangle*).

\[
\text{Prob (transition|} t\text{)} = \frac{1}{4} - \frac{1}{2} \exp \left( -\frac{R+\frac{1}{2}}{R+1} t \right) + \frac{1}{4} \exp \left( -\frac{2}{R+1} t \right)
\]

\[
\text{Prob (transversion|} t\text{)} = \frac{1}{2} - \frac{1}{2} \exp \left( -\frac{2}{R+1} t \right).
\]

(1)

(the transversion probability is the sum of the probabilities of both kinds of transversions).
Fractions of transitions and transversions expected in different amounts of branch length under the K2P model, for $T_s/T_n = 10$
Fractions of transitions and transversions expected in different amounts of branch length under the K2P model, for $T_s/T_n = 2$
Other commonly used models include:

Two models that specify the equilibrium base frequencies (you provide the frequencies $\pi_A, \pi_C, \pi_G, \pi_T$ and they are set up to have an equilibrium which achieves them), and also let you control the transition/transversion ratio:

The **Hasegawa-Kishino-Yano (1985) model**:

<table>
<thead>
<tr>
<th>to :</th>
<th>$A$</th>
<th>$G$</th>
<th>$C$</th>
<th>$T$</th>
</tr>
</thead>
<tbody>
<tr>
<td>from :</td>
<td>$A$</td>
<td>$-\alpha\pi_A + \beta\pi_A$</td>
<td>$\alpha\pi_G + \beta\pi_G$</td>
<td>$\alpha\pi_C$</td>
</tr>
<tr>
<td>$G$</td>
<td>$\alpha\pi_A + \beta\pi_A$</td>
<td>$-\alpha\pi_C$</td>
<td>$\alpha\pi_G$</td>
<td>$\alpha\pi_T$</td>
</tr>
<tr>
<td>$C$</td>
<td>$\alpha\pi_A$</td>
<td>$\alpha\pi_G$</td>
<td>$-\alpha\pi_C + \alpha\pi_T + \beta\pi_T$</td>
<td></td>
</tr>
<tr>
<td>$T$</td>
<td>$\alpha\pi_A$</td>
<td>$\alpha\pi_G$</td>
<td>$\alpha\pi_C + \beta\pi_C$</td>
<td>$-$</td>
</tr>
</tbody>
</table>

My **F84** model:

<table>
<thead>
<tr>
<th>to :</th>
<th>$A$</th>
<th>$G$</th>
<th>$C$</th>
<th>$T$</th>
</tr>
</thead>
<tbody>
<tr>
<td>from :</td>
<td>$A$</td>
<td>$-\alpha\pi_A + \beta\pi_A$</td>
<td>$\alpha\pi_G + \beta\pi_G\frac{\pi_R}{\pi_Y}$</td>
<td>$\alpha\pi_C$</td>
</tr>
<tr>
<td>$G$</td>
<td>$\alpha\pi_A + \beta\pi_A\frac{\pi_R}{\pi_Y}$</td>
<td>$-\alpha\pi_C$</td>
<td>$\alpha\pi_G$</td>
<td>$\alpha\pi_T$</td>
</tr>
<tr>
<td>$C$</td>
<td>$\alpha\pi_A$</td>
<td>$\alpha\pi_G$</td>
<td>$-\alpha\pi_T + \beta\pi_T\frac{\pi_Y}{\pi_R}$</td>
<td></td>
</tr>
<tr>
<td>$T$</td>
<td>$\alpha\pi_A$</td>
<td>$\alpha\pi_G$</td>
<td>$\alpha\pi_C + \beta\pi_C\frac{\pi_Y}{\pi_R}$</td>
<td>$-$</td>
</tr>
</tbody>
</table>

where $\pi_R = \pi_A + \pi_G$ and $\pi_Y = \pi_C + \pi_T$ (The equilibrium frequencies of purines and pyrimidines)

Both of these models have formulas for the transition probabilities, and both are subcases of a slightly more general class of models, the **Tamura-Nei model** (1993).
There is also the **General Time-Reversible model (GTR)**, which maintains “detailed balance” so that the probability of starting at (say) A and ending at (say) T in evolution is the same as the probability of starting at T and ending at A:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>G</th>
<th>C</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>from:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>−</td>
<td>απ&lt;sub&gt;G&lt;/sub&gt;</td>
<td>βπ&lt;sub&gt;C&lt;/sub&gt;</td>
<td>γπ&lt;sub&gt;T&lt;/sub&gt;</td>
</tr>
<tr>
<td>G</td>
<td>απ&lt;sub&gt;A&lt;/sub&gt;</td>
<td>−</td>
<td>δπ&lt;sub&gt;C&lt;/sub&gt;</td>
<td>επ&lt;sub&gt;T&lt;/sub&gt;</td>
</tr>
<tr>
<td>C</td>
<td>βπ&lt;sub&gt;A&lt;/sub&gt;</td>
<td>δπ&lt;sub&gt;G&lt;/sub&gt;</td>
<td>−</td>
<td>υπ&lt;sub&gt;T&lt;/sub&gt;</td>
</tr>
<tr>
<td>T</td>
<td>γπ&lt;sub&gt;A&lt;/sub&gt;</td>
<td>επ&lt;sub&gt;G&lt;/sub&gt;</td>
<td>υπ&lt;sub&gt;C&lt;/sub&gt;</td>
<td>−</td>
</tr>
</tbody>
</table>

And there is of course the **general 12-parameter model** which has arbitrary rates for each of the 12 possible changes (from each of the 4 nucleotides to each of the 3 others). Neither of these has formulas for the transition probabilities, but those can be done numerically.
There are many other models, but these are the most widely-used ones.

Here is a general scheme of which models are subcases of which other ones:

- General 12-parameter model (12)
- General time-reversible model (9)
- Tamura–Nei (6)
  - HKY (5)
  - F84 (5)
    - Kimura K2P (2)
      - Jukes–Cantor (1)
Gamma distributions with mean 1 and different coefficients of variation (standard deviation / mean). $\alpha = 1/CV^2$ is the “shape parameter” of the Gamma distribution.
References:


